International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(15s) 2026 pp. 915-921
https://doi.org/10.70917/ijcisim-2026-4491
Article

Enhancing Heart Disease Prediction Through The Hybrid Random Forest–Gradient Boosting-Logistic Regression Model (HRFGLM): A Data-Driven Predictive Framework

Asha Dilipkumar Jariwala[1], Hemangini Patel[2]

[1] Department of Computer Science, Vanita Vishram Women's University, Surat, Gujarat, India.
Email: ahnagoriya@gmail.com
[2] Department of Computer Science, Vanita Vishram Women's University, Surat, Gujarat, India.
Email: hemangini.patel@vvwusurat.ac.in

Abstract: — heart disease is currently one of the most significant issues facing the world. One of the most significant illnesses affecting blood vessels and the heart is cardiovascular disease. The toll of death from cardiovascular disease, which is mostly caused by a lack of early disease detection, will be greatly decreased if the risk is predicted beforehand. Anticipating cardiac disease could be a significant medical breakthrough because it is so prevalent. Machine learning approaches anticipate the disease based on the severity of the patient's side effects due to the growing amount of data in the healthcare industry. This study suggests a methodology for early cardiovascular disease prediction using various machine learning techniques for various prediction objectives. Nevertheless, a number of these methods might be enhanced, such as inadequate accuracy. In our research, we have taken the cardicascular diseases dataset and implemented a few models, which include Gradient Boost, Random Forests, and Linear Regression classifiers getting 75.19% accuracy. This work has the advantage of using machine-learning techniques to improve coronary heart disease prediction performance.
Keywords: — Cardiovascular Disease Prediction, heart disease prediction, Machine Learning algorithms, Feature selection, overfitting issues

1. INTRODUCTION

Nowadays, Heart disease has become a major worldwide health issue. [1] cardiovascular diseases (CVDs) refer to a range of disorders that impact the blood and heart vessels. Cardiovascular disorders are diseases of the blood vessels. The World Health Organization (WHO) claims that cardiovascular disease (CVD) is the most common cause of mortality worldwide, causing more fatalities each year than all other causes combined.[2] In order to protect afflicted people, prompt diagnosis is essential. Furthermore, food and lifestyle choices have a big impact on a person's health, therefore it's critical to examine a patient's medical history in order to forecast and identify HD early [3].

By pumping blood through an intricate network of blood veins, the heart keeps the human body alive. Circulation of blood removes air and other metabolic waste while supplying oxygen and other vital nutrients to all bodily tissues. The fact that decreased heart efficiency immediately reduces the functional capacity of all other organs, including the brain, kidneys, and lungs, emphasizes the critical role the heart plays in maintaining the human body. Additionally, the heart is essential for controlling blood pressure, body temperature, and the balance of biochemicals and hormones in the blood. Heart failure, hypertensive heart disease, myocardial infarction (heart attack), poor dietary habits, inactivity, smoking, excessive alcohol use, obesity, and prolonged stress are all associated with coronary artery disease (CAD) [4].Heart conditions encompass several serious disorders affecting the heart's function, as illustrated in the fig.1[1]

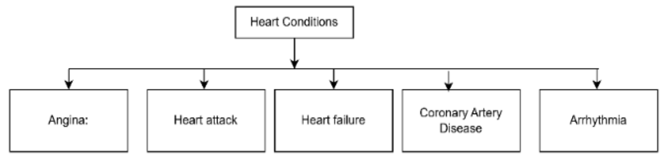

Fig. 1. Common Heart Conditions

Disorders that impact heart function are included under heart conditions. Angina (reduced blood flow), heart attack (blocked arteries), heart failure (poor pumping), coronary artery disease (plaque accumulation), and arrhythmia (irregular heartbeat) are the main forms. It's critical to identify symptoms like dyspnea and chest pain as soon as possible [1]. Fitness and health bands are just two examples of the many advancements in the medical industry. Additionally, tools like CT scans and electrocardiograms aid in the identification of coronary heart disease [5]. Predictive models were constructed using a range of methods for supervised learning such as Logistic regression, Gradian boost and Random Forest, a user-friendly platform popular for machine learning applications, was used to implement and assess these models. This study investigates the use of machine learning algorithms for heart disease prediction using a cardiovascular diseases dataset that contains 12 features, including blood pressure, cholesterol, gender, age and other health-related indicators. The purpose of this work is to support the development of systems by Predicting heart disease early with machine learning accurate diagnoses of heart diseases. The rest of this paper is organized as follows: Section 2 Describe data and methods Section 3 describes the literature review; Section 4 presents the methodology; Section 5 is the discussion; Section 5 presents the result analysis; Section 6 Conclusion.

2. DATA AND METHODS

We used Kaggle cardiovascular dataset In Table 1, The dataset contains 70000 rows and 12 columns, about patient characteristics such as gender, height, age weight, systolic blood pressure, smoke, glucose, diastolic blood pressure alcohol, active and cholesterol. In Fig. 2.it shows the overall flow of methodology. In data collection we collected the data from Kaggle site. Following that, the final cleared data was 68517. Out of 68517, 33192 had cardiac. There were 11946 males and 21966 females in that. The minimum age is 39 and the maximum age is 45. After that did feature engineering and add more four columns Pulse pressure, MAP, bp category and BMI. Afterward divided the data in 20% testing data and 80% Training Data. We used Hybrid model combination of (Random Forest, Gradian Boost, Logistic Regression).

Machine behavior that happens when training data is accurately predicted by the machine learning model, but new data is not. Data scientists first train machine learning models on a known data set before using them to make predictions. The model then attempts to forecast results for fresh data sets using this information.

Attribute	**Description**
Age	Objective Feature | age | int (days)
Height	Objective Feature | height | int (cm)
Weight	Objective Feature | weight | float (kg)
Gender	Objective Feature | gender | categorical code
Systolic blood pressure	Examination Feature | ap_hi | int
Diastolic blood pressure	Examination Feature | ap_lo | int
Cholesterol	Examination Feature | 1: normal cholesterol; 2: above normal; and 3: significantly above normal
Glucose	Examination Feature | Gluc 1: typical 2: above average 3: significantly over average

Smoking	Subjective Feature | smoke | binary
Alcohol	Subjective Feature | alco | binary
Physical	Subjective Feature | active | binary
Cardiovascular disease's presence or absence	Target Variable | cardio | binary

TABLE 1. Dataset

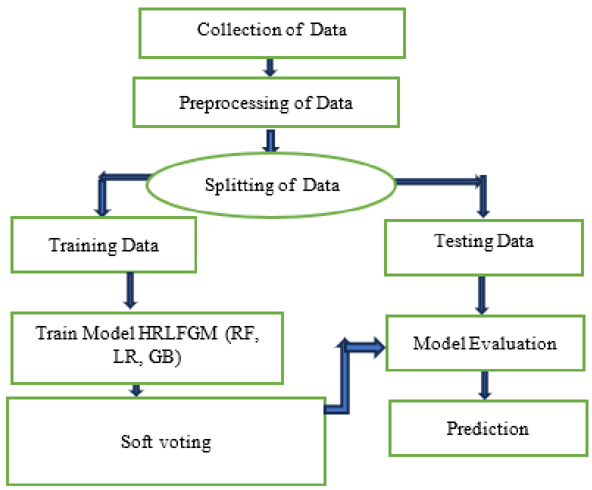

Fig. 2. Flowchart of Methodology

3. LITERATURE REVIEW

It shows the actual outcome of the used Gaussian SVM kernel classifier accuracy 86.3% [1]. It used, the ontology classifier has the best accuracy (75.5%), followed by the Decision Tree (73.1%) and LR (72.1%) [2]. Hybrid extreme ML learning model. HEMLM method combines the advantages of logistic regression (LR), random layers, and multi-layer perceptron's (MLP). The HEMLM algorithm has exceptional efficiency and performance. 87.14% precision [3]. A suggested stacking method uses LR as the meta-classifier and integrates the results of RF, KNN, and SVM. Using the Kaggle CVD dataset, which has 70,000 patients' data along with 12 attributes, this approach produced 75.1% accuracy [5]. KNN, Random Forest, Naive Bayes, and Logistic Regression were among the ML methods evaluated. On the Kaggle CVD dataset, which included 70,000 patients and 12 attributes, their best model produced 70% accuracy [6]. It achieved a 72.77% accuracy rate in predicting cardiac disease using a Decision Tree classifier. The Kaggle cardiovascular dataset, which includes 70,000 patients and 12 attributes, was also used in the research [8]. A RF classifier and a repeated random validation procedure yielded the greatest accuracy of 89.01%.The UCI cardiovascular dataset, which comprises 303 patients and 14 characteristics, was used in their tests [9]. Using a neural network for holdout cross-validation, they obtained an accuracy of 71.82% on the Kaggle cardiovascular sickness dataset, which included 70,000 patients and 12 attributes [10]. It suggested utilizing the UCI cardiac to forecast the risk of heart disease utilizing a HRFLM model. Their method produced an accuracy of 88.7%, indicating that prediction effectiveness was increased when RF and a Linear Model were combined [11].HRFLM (Random Forest + Logistic Regression) was used to predict cardiovascular disease using the Cleveland dataset. Their method achieved maximum 98.54% accuracy, Beyond traditional ML methods in terms of interpretability [12]. Using the UCI Cleveland dataset, it created a hybrid machine learning model that predicts heart disease using Random Forest, Logistic Regression, and Naïve Bayes. They reported 88.7% accuracy with an F1-score of 0.87 and observed enhanced robustness due to the hybrid approach [13]. By applying HRFLM to the UCI dataset and obtaining an accuracy of 87.7%, it showed improved prediction with effective training time [14]. Using HRFLM (Random Forest + Linear Regression) on the Cleveland dataset, it attained 88% accuracy. They found that the hybrid model enhanced generalization and precision [15].

4. CLASSIFICATION

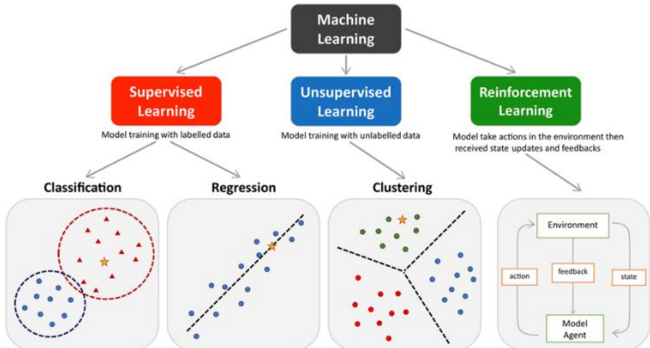

Fig3. Types of Machine Learning [16]

In Fig 3 shows the primary categories of machine learning. The two primary methods are clustering for unsupervised learning and classification and regression for supervised learning. There are a lot of different things that can be done. By interacting with the environment, reinforcement learning improves the model's performance. The training data is represented by coloured triangles and dots. The new data that the trained model is able to anticipate is represented by yellow stars.

Random Forest

Straightforward and versatile Random Forest supervised machine learning approach can be used to address regression and classification problems. It will use ensemble learning methods to construct trees. Using the data samples, the RF approach builds several decision trees, gathers predictions from each tree, and then uses majority voting to choose the best option.

Logistic Regression

The sigmoid function, which receives input as independent variables and yields a probability number from 0 and 1, is used in logistic regression for binary classification.

Gradient boosting

A powerful ensemble learning technique in machine learning for both regression and classification applications is a Gradient Boosting (GB). By gradually merging multiple "weak" models—typically shallow decision trees—it creates a single "strong" prediction model. In order to enhance performance, it builds models one after the other with a focus on correcting errors from previous models. There are a lot of different kinds of things that can happen.

Hybrid Ensemble Model (HRFGLM). *Fig 4* shows that combines the strengths of Random Forest, Gradient Boosting and Logistic Regression utilize a soft voting classifier-it uses each model's probability. Because the weights in our model are RF=2, LR=1, and GB=4, Gradient Boosting has a greater influence on the outcome. Soft voting makes use of both final class labels and the anticipated probability from each model.

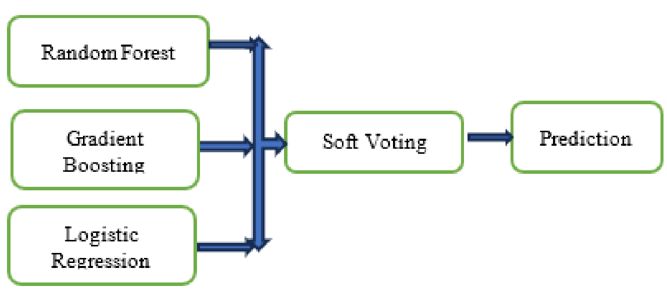

Fig 4. HRFGLM Model

5. EXPERIMENTAL RESULT

Fig 5.shows the confusion Matrix. A confusion matrix shows how well your model classified the two classes: 0=No Heart Problem ,1 = Heart Problem.True positive(TP) values These are real patients without cardiac disease who were accurately predicted to have a score of 0 which is 2776. False Positive (FP)means Actual class = 0, but predicted 1 value is 685.False Negative (FN) means Actual class = 1, but predicted 0 value is 1015. True Positive (TP) means Actual class = 1, predicted 1 value is 2376. *Fig 6*.Classification reports included Precision ,recall,accuracy and F1-score

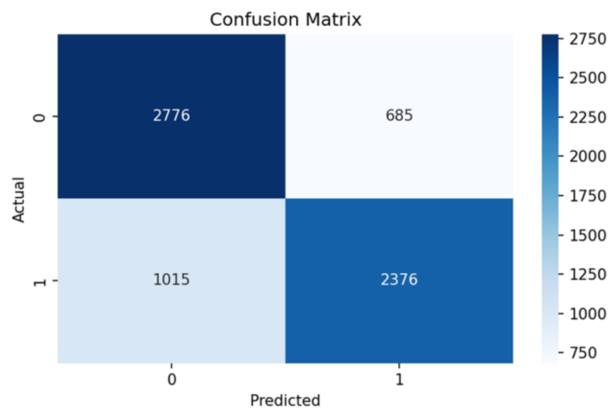

Fig. 5 Confusion matrix

Class	**Precision**	**Recall**	**F1-Score**	**Support**
No Cardio	0.73	0.80	0.77	3461
Cardio	0.78	0.70	0.74	3391
Accuracy			**0.75**	6852
Macro Avg	0.75	0.75	0.75	6852
Weighted Avg	0.75	0.75	0.75	6852

TABLE 2. Classification Report — HRFGLM

Table 2 defines classification report of the HRFGLM model*Fig 6*.shows Area Under the Receiver Operating Characteristic Curve, or AUC-ROC curve. The model uses RFE-selected features to differentiate between positive and unfavorable circumstances.It assesses how well a classification model distinguishes between classes. The ROC curve's smooth, continually increasing shape indicates stable and reliable predictions..

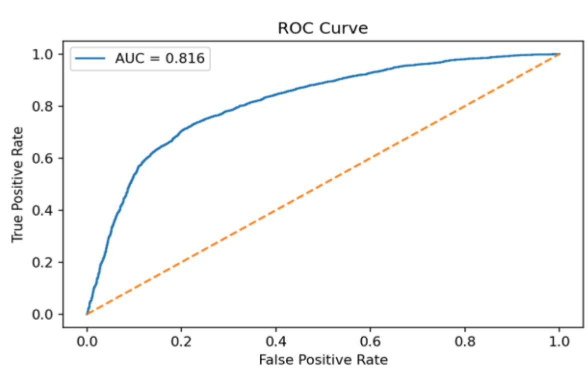

Fig. 6. AUC-ROC Curve for HRLFM with REF

In *Fig 7.* Strong positive correlations between blood pressure variables (ap_hi, ap_lo, MAP, pulse_pressure, bp_category) are highlighted in the heatmap, which displays the link between heart-disease-related aspects.

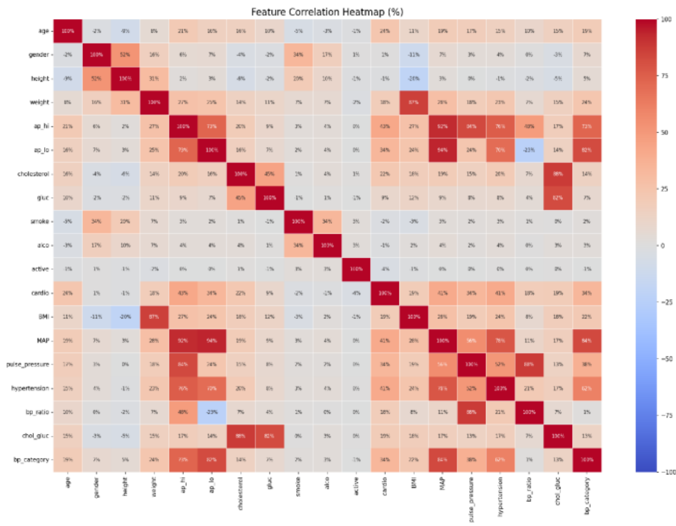

Fig. 7. Feature Correlation Heatmap

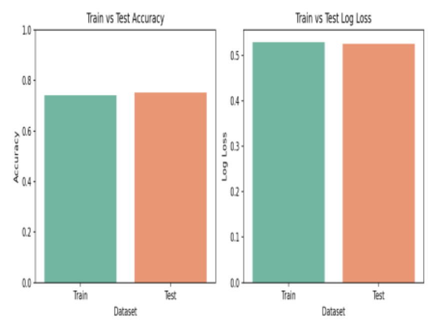

Fig.8 overfitting Analysis

The overfitting analysis *Fig.8* HRFGLM model generalizes well, as the training and testing accuracies *Gap* is −1.01%. Since the gap is negative and very small, there is no severe overfitting. *Table 3*. shows Overfitting range Thresholds are supported by practical ML evaluation standards

Range	value
0.00–0.03	(No Overfitting)
0.03–0.07	(Mild Overfitting)
>0.07	(Strong Overfitting)

Table 3 Overfitting range Thresholds are supported by practical ML evaluation standards

6. CONCLUSION

The HRFGLM model gave a good and stable performance for cardiovascular disease prediction. It achieved 75.19% testing accuracy and 0.8160 ROC AUC. RFC+SVC model accuracy is 74.70.KN+RF+SVC model accuracy is 75.10. DT +NB model accuracy is 73.70. HRLFM model accuracy is 73.65 HRFGLM is not facing overfitting issues because training and testing accuracy are close and also achieved the curse of dimensionality. RFE selected 19 important features, and the hybrid soft voting approach helped combine the strengths of RF, LR, and Gradient Boosting. Overall, HRFGLM is reputable model for this dataset.

References

1. Kotia, A. S. S., Rastogi, M., & Bhongade, R. A. (2023). Use of machine learning techniques for effective prediction of heart disease. CARDIOMETRY, 315-321.
2. Massari, H. E., Gherabi, N., Mhammedi, S., Ghandi, H., Bahaj, M., & Naqvi, M. R. (2024). The impact of ontology on the prediction of cardiovascular disease compared to machine learning algorithms. arXiv preprint arXiv:2405.20414.
3. Ahmed, A. M., Bataineh, B., Shakah, G., Al Enany, M. O., Aboghazalah, M. M., & Khattab, M. M. (2025). A hybrid extreme machine learning model for predicting heart disease. Bulletin of Electrical Engineering and Informatics, 14(5), 4125-4137.
4. Banerjee, T., & Paçal, İ. (2025). A systematic review of machine learning in heart disease prediction. Turkish Journal of Biology, 49(5), 600-634.
5. Shorewala, V. Early detection of coronary heart disease using ensemble techniques. Inform. Med. Unlocked 2021, 26, 100655.
6. "Predicting Cardiovascular Disease Using Machine Learning Techniques 1Harichselvam C, 2Dr. Jerald F, 3Dr. T.V. Ananthan, 4Harishwar J, 5Maheswari
7. .Maiga, J.; Hungilo, G.G.; Pranowo. Comparison of Machine Learning Models in Prediction of Cardiovascular Disease Using Health Record Data. In Proceedings of the 2019 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), Jakarta, Indonesia, 24–25 October 2019; pp. 45–48.
8. .Waigi, R.; Choudhary, S.; Fulzele, P.; Mishra, G. Predicting the risk of heart disease using advanced machine learning approach. Eur. J. Mol. Clin. Med. 2020, 7, 1638–1645.
9. .Ouf, S.; ElSeddawy, A.I.B. A proposed paradigm for intelligent heart disease prediction system using data mining techniques. J. Southwest Jiaotong Univ. 2021, 56, 220–240.
10. .Khan, I.H.; Mondal, M.R.H. Data-Driven Diagnosis of Heart Disease. Int. J. Comput. Appl. 2020, 176, 46–54.
11. Sivaji, U., Rao, N. K., Srivani, C., Sree, T., & Singh, M. (2021). A Hybrid Random Forest Linear Model approach to predict the heart disease. Annals of the Romanian Society for Cell Biology, 25(6), 7810-7814.
12. Sathwik, A. S., Naseeba, B., & Challa, N. P. (2023, July). Cardiovascular Disease Prediction Using Hybrid-Random-Forest-Linear-Model (HRFLM). In 2023 IEEE World Conference on Applied Intelligence and Computing (AIC) (pp. 192-197). IEEE
13. Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. IEEE access, 7, 81542-81554.
14. Sweatha, M., Ramya, S., Icewariiya, S., & Kalaivani, V. (2022). Heart disease prediction using HRFLM. International Journal of Advances in Engineering and Management (IJAEM), 4(6), 1619-1623. https://doi.org/10.35629/5252-040616191623
15. Khan, Z., Sikandar, G., & Anwar, S. (2023). Heart disease prediction using hybrid random forest and linear model. International Journal of Emerging Engineering and Technology (IJEET, 2(1), 6–12.
16. Peng, J., Jury, E. C., Dönnes, P., & Ciurtin, C. (2021). Machine learning techniques for personalised medicine approaches in immune-mediated chronic inflammatory diseases: applications and challenges. Frontiers in pharmacology, 12, 720694.